DOCUMENT RESUME

ED 464 963 TM 033 894

AUTHOR Lee, Yong-Won; Breland, Hunter; Muraki, Eiji
TITLE Comparability of TOEFL CBT Essay Prompts for Different Native Language Groups.
INSTITUTION Educational Testing Service, Princeton, NJ.
PUB DATE 2002-04-04
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Comparative Analysis; *Computer Assisted Testing; English (Second Language); *Essay Tests; *Item Bias; Language Minorities; Second Language Learning
IDENTIFIERS Europeans; Logistic Regression; Nonnative Speakers; *Test of English as a Foreign Language; *Writing Prompts

ABSTRACT

Since the writing section of the Test of English as a Foreign Language (TOEFL) computer based test (CBT) is a single-prompt essay test, it is very important to ensure that each prompt is as fair as possible to any subgroups of examinees, such as those with different native language backgrounds. A particular topic of interest in this study is the question of whether the examinees of non-Indo-European native language backgrounds can be unduly disadvantaged on some essay prompts compared to those of European language backgrounds. A total of 81 TOEFL CBT writing prompts (262,034 essays) introduced from July 1998 through July 2000 were examined for essay prompt differential item functioning (DIF). The English language ability (ELA) variable was created by summing the standardized TOEFL Reading, Listening, and Structure scores and used to match examinees in the East Asian (Chinese, Japanese, and Korean) and European (German, French, and Spanish) native language groups. A three-step logistic regression modeling procedure for polytomous items was used to examine both the uniform and nonuniform effects of the native languages on the essay scores. Mean ELA, raw essay, and residual-based effect sizes were also computed for each prompt. Although approximately one-third of the prompts analyzed had statistically significant group differences ($p=0.05$), the R squared effect size was too small for any prompt to be classified as having a group effect. An appendix describes the derivation of the logistic regression model used. (Contains 5 tables, 4 figures, and 29 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 963

Comparability of TOEFL CBT Essay Prompts for Different Native Language Groups

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Yong-Won Lee

2A

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee

Hunter Breland

Educational Testing Service (ETS)

and

Eiji Muraki

American College Testing (ACT)

Paper presented at the annual meeting of National Council on Measurement in Education (NCME) held between April 1-5 in New Orleans, LA.

TM033894

Unpublished Work © 2002 by Educational Testing Service. All Rights Reserved.

BEST COPY AVAILABLE

Abstract

Since the writing section of TOEFL CBT is a single-prompt essay test, it is very important to ensure that each prompt is as fair as possible to any subgroups of examinees (*e.g.*, examinees with different native language backgrounds). A particular topic of interest in this study is the question of whether the examinees of non-Indo-European native language backgrounds can be unduly disadvantaged on some essay prompts compared to those of European language backgrounds. A total of 81 TOEFL-CBT writing prompts introduced from July 1998 through July of 2000 were examined for essay prompt DIF (differential item functioning). The English language ability (ELA) variable was created by summing the standardized TOEFL Reading, Listening, and Structure scores and used to match examinees in the East Asian (Chinese, Japanese, and Korean) and European (German, French, and Spanish) native language groups. A three-step logistic regression modeling procedure for polytomous items was used to examine both the uniform and non-uniform effects of the native languages on the essay scores. Mean ELA, raw essay, and residual-based effect sizes were also computed for each prompt. Although approximately one-third of the prompts analyzed had statistically significant group differences (p=0.05), the R^2 effect size was far too small for any prompt to be classified as having a group effect.

ETS Educational Testing Service

Introduction

CBT administrations of the Test of English as a Foreign Language (TOEFL) began in the summer of 1998. These administrations included computer-based multiple-choice (MC) tests of reading, listening, structure (a multiple-choice test of grammar and sentence structure), and writing. The prompts for the CBT essay section are selected for each examinee from a pool of prompts in a near-random manner, and as a result all examinees do not receive the same prompt. Since only one essay prompt is administered per examinee in TOEFL CBT, it becomes a very important issue to ensure that each writing prompt is as fair as possible to any subgroups of examinees who take the same writing prompt from the pool of prompts. In addition, it would be equally important to understand the causes of such group-related effect on essay scores and to introduce procedures to minimize any potential biases. From the score interpretation perspective, when the prompt incomparability is caused by serious prompt bias against a certain subgroup (s) of examinees, it could potentially distort the meaning of the essay score for different examinee subgroups and become a potential threat to test score validity (Sheppard, 1982).

For this reason, a considerable amount of research effort has been directed toward investigating the comparability of TOEFL CBT writing prompts for different test-taker subgroups, such as different gender and entry mode groups (Breland, Muraki, Lee, 2001; Breland, Muraki, Lee, Najarian & Beyer, 2000; Gentile & Riazantseva, 2001; Wolfe & Minalo, 2000). These studies have found that there might be a small-to-medium impact of gender and entry modes on essay scores and provided some useful information for subsequent prompt review, revision, and retirement to minimize the potential impact. One important group variable that has yet to be examined fully for the TOEFL CBT writing test is native language backgrounds of test-takers, as the AERA/APA/NCME (1999) standards also suggest. It has been suggested that the native language backgrounds of the examinees could have impact on their performance on second/foreign language tests. Of particular interest has been the question of whether the examinees of non-Indo-European language backgrounds can be unduly disadvantaged on the EFL/ESL (English as a foreign/second language) tests compared to those of Indo-European language backgrounds due to the relatively greater dissimilarity between their native languages and the English language.

The impact of examinees' native languages on their performance on second or foreign language tests has long been an important issue for test developers and users (Ginther & Stevens, 1998; Oltman, Stricker & Barrows, 1988; Swinton & Powers, 1980). Differential item function (DIF) studies in particular have also investigated test-takers' differential performance on the test that could be attributable to their native languages (Angoff, 1989; Chen & Henning, 1985; Kim, 2001; Ryan & Bachman, 1992; Sasaki, 1991) in the ESL/EFL (English as a Second/Foreign Language) context. However, these previous DIF studies in language assessment have some major limitations: (1) item responses studied were confined to dichotomously-scored multiple-choice (MC) language tests including vocabulary, grammar, and listening, and reading comprehension; (2) and only uniform DIF was the main focus of the study in most of these studies; (3) all of these studies dealt with situations where the internal matching criterion is usually available for

the studied items. Among these studies, Kim's (2001) DIF study on the pre-revision version of the Test of Spoken English (TSE) might be an exception to limitations (1) and (2), but, even for her study, an internal matching criterion was available. Clearly, there has been a lack of DIF research on essay tests that are scored polytomously, where no reliable, internal matching criterion is usually available.

The research study reported in the present paper attempts to move one step beyond such limitations and to examine the essay prompt DIF for different native language groups. To select the most feasible scheme for DIF investigation in this study, the characteristics of matching criteria, items, and kinds of analyses employed had to be carefully considered at the beginning of this study. Since there was no internal matching criterion, an externally available criterion in the same test battery was used. In addition, a DIF method had to be chosen for the study that could examine both the uniform and non-uniform effects on examinees. The method selected was *logistic regression* of essay scores on the matching variable. A more detailed rationale for the methodology adopted for this study follows:

One of the most important challenges in this project was to find an appropriate variable to use for matching examinees of two different subgroups on their English writing ability. By definition, DIF occurs when the examinees of equal ability but from the different groups do not have the same probability of success on an item (Angoff, 1993; Hambleton, Swaminathan & Rogers, 1991). Because the TOEFL CBT writing section is an essay test made up of a single prompt, no internal matching criterion for writing ability is available (Potenza & Dorans, 1995). The only information available was that provided by the three multiple-choice sections of TOEFL CBT (listening, structure, and reading). Since the structure section and the essay scores constitute the structure/written expression (SWE) section score in an operational sense, one might argue that the structure section scores alone could be a more valid matching variable for the writing ability. However, the structure items may not be measuring exactly the same construct measured by the essay test, either (DeMauro, 1992). The correlation between the essay rating and the structure score is not significantly larger than that between the essay and other two multiple-choice section scores. Rather, when these three sections were standardized and combined, the correlation between the essay scores and the matching criterion was maximized. For these reasons, a decision was made to create a matching variable by summing the standardized scaled scores from the three multiple-choice sections on the assumption that, if examinees had high English language ability measured by the three sections of the test as a whole, they would be also expected to perform well overall on the essay scores, and vice versa.

A second challenge related to the nature of the essay scores was that examinees' essays for each prompt are scored according to a multi-point scoring rubric by two independent raters. The final two raters' ratings are averaged and rounded off for score reporting purposes. The essay scores are thus discrete and bounded between 1 and 6. This poses a challenge for DIF analysis because most of the popular DIF methodology has been developed with a focus on dichotomous items. A DIF method was required that could handle polytomously-scored items. There have been various methods of investigating DIF extended for polytomous items including logistic regression (French & Miller, 1996), polytomous IRT (item response theory) (Muraki, 1999; Wainer, Sireci, & Thissen, 1991),

the Mantel and Hanszel technique (Zwick, Donoghue, & Grima, 1993), the SIBTEST procedure (Chang, Mazzeo, & Roussos, 1995), the standardization method (Dorans & Schmitt, 1991), and logistic discriminant function analysis (Miller & Spray, 1993). Among these methods, however, DIF methods requiring an internal criterion (e.g., polytomous IRT and polytomous SIBTEST) were not feasible for this study. Moreover, methods such as the Mantel and Hanszel technique, SIBTEST, and the standardization methods, may not be appropriate for detecting the non-uniform DIF (Penfield & Lam, 2000), unless a special modification is made (Clauser & Mazor, 1998).

A third challenge resulted from the need to examine non-uniform as well as uniform DIF, since this might be critical in the investigation of the impact of native languages on the TOEFL CBT essay scores. There has been a considerable amount of research on the effect of test-takers' native language backgrounds on the test dimensionality of the EFL/ESL test or on the different factor structure of the same tests across the different levels of language proficiency (Ginther & Stevens, 1998; Oltman, Stricker & Barrows, 1988; Swinton & Powers, 1980). Some of these studies have suggested that the interpretation of TOEFL section scores depends on the examinees' overall level of proficiency, with more differentiated factors for TOEFL sections for low-proficiency examinees, but less distinct constructs for high-scoring examinees. Since the matching variable is from the MC sections of the listening, structure, and reading, an argument can be made that the native language backgrounds might have differential impact on essay scores depending upon the English ability levels of the examinees. In addition, the presence of strong non-uniform DIF may be very closely related to task validity that extend beyond the realm of DIF, as Penfield and Lam (2000) have pointed out. Both the IRT and the logistic regression can be very effective in detecting non-uniform DIF, but the IRT method needs an internal matching criterion. A clear advantage of the logistic regression for non-uniform DIF is that it is close to IRT in form but not constrained by sample size and model-data fit as much as in IRT methods.

The principal objective of this study was to investigate the comparability of TOEFL CBT writing prompts for examinees of different language backgrounds, especially focusing on the East Asian (Chinese, Japanese, and Korean) versus European (German, French, and Spanish) native language groups as focal and reference groups, respectively. More specifically, this study is designed to assess the performance of examinees in the Asian and European language groups after they were matched on English language ability, as defined in this study. Methodologically, this study is primarily concerned with procedures for examining non-uniform and uniform effects of an examinee group variable in an essay test, where examinees take usually a single essay prompt that is scored polytomously.

Methods

Sample

The sample of data analyzed consisted of TOEFL CBT essay data collected through

August 2000 administrations for three European (*i.e.*, French, German, and Spanish) and three East Asian (*i.e.*, Chinese, Japanese, and Korean) language groups. In total, 262,034 essays were written on 87 different topics between July of 1998 and July of 2000. Six prompts with insufficient data for the focal and reference groups were dropped from the analysis. Of the 254,435 examinees and 81 prompts included in the analysis, a total of 121,494 examinees were native speakers of three European languages and 132, 941 were native speakers of three East Asian languages. Among the three European languages, Spanish was the largest group (n=66,282), followed by French (n=28,007) and German (n=27,205). Chinese was the largest group (n=52112) among the Asian languages, followed by Japanese (n=43,666) and Korean (n=37,163).

Instruments

Data analyzed included scores on the reading, listening, structure, and writing subtests of the TOEFL CBT. The listening and structure tests are adaptive, the reading test is linear, and the writing test score is basically the average of the final two reader ratings. If the first two readers' ratings differ by more than one point, a third rater adjudicates the rating discrepancy. The two closest ratings are averaged and rounded off to the nearest 0.5 (*e.g.*, 1.0, 1.5,….5.5, 6.0). The English Language Ability (ELA) variable was created by summing the three standardized multiple-choice section scores (*e.g.*, reading, listening, and structure sections).

Data Analysis

Logistic regression analysis (Hosmer & Lemeshow, 1989) has been used mainly to detect dichotomous DIF by specifying separate equations for the focal and reference subgroups of examinees (Swaminathan & Rogers, 1990). French and Miller (1996) have demonstrated that this procedure can be extended for polytomous DIF as well. In this study, the logistic regression procedure by French and Miller (1996) is extended further to be used for TOEFL CBT essay prompts DIF. Logistic regression has also two main advantages over linear regression. The first is that the dependent variable does not have to be continuous, unbounded, and measured on an interval or ratio scale. The second is that it does not require a linear relationship between the dependent and independent variables. In the case of TOEFL data, the dependent variable (the essay score) is discrete and bounded between 1 and 6. In addition, there is evidence of non-linearity in the relationship between essay scores and the independent variable (ELA) for the various group memberships. Violation of both of these assumptions indicated that a logistic regression model would be most appropriate for the study.

The logistic regression method employed in this study was the "proportional odd-ratio model" that is also implemented in the SAS logistic procedure (SAS Institute, 1990). A 3-step modeling process based on logistic regression (Zumbo, 1999) was used as a main method of analysis along with a residual-based procedure devised for this study. Polytomous essay scores were dichotomized into 10 binary variables according to the

cumulative-logit dichotomization scheme (See Appendix A for more details). The 10 dichotomized essay variables were simultaneously regressed on examinees' ELA scores, the native language dummy group variable (European =0; Asian 1), and the ability-by-group interaction variable in a step-by-step fashion. Equal slopes were assumed for the all of the 10 dichotomized variables from the same prompt. More specifically, the ordinal logistic regression analysis was conducted in the following three steps: In Step 1, the matching variable or the conditioning variable (*i.e.*, ELA scores) was entered into the regression equation for all the dichotomized responses (i), as in $g_i(x, D) = \beta_{0i} + \beta_1 x$; In Step 2, the group membership (*i.e.*, European vs. Asian) variable was entered ($g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m$); In Step 3, the interaction term (*i.e.*, English language ability-by-group) was finally added ($g_i(x, D) = \beta_{0i} + \beta_1 x + \beta_2 D_m + \beta_3 x D_m$). The three nested models in steps 1-3 can be fitted to the data and compared in terms of model-data fit (expressed in terms of χ^2 statistics) and of the size of R^2 coefficients.

To gauge the amount of the group differences (if any), three different kinds of the effects sizes from the logistic regression were used in this study: (1) the residual-based effect size; (2) R^2 combined with p-values for the χ^2 test and slope parameters; and (3) the group-specific expected score curves. Before the full 3-step modeling process was begun, expected essay scores, residual scores, and the residual-based effect sizes were computed for all the prompts by using only the matching variable (*i.e.*, ELA scores) in the regression model. Expected essay scores for individual examinee's ELA scores were computed from the step-one model ($g_i(x, D) = \beta_{0i} + \beta_1 x$). Residual scores were obtained for each individual examinee by subtracting their ELA- predicted essay scores from their raw essay scores, and these residual scores were averaged separately for each language group on each prompt. The residual-based effect sizes were computed by dividing the mean residual score difference between the two groups by the pooled standard deviation of the raw essay scores for both language groups. The residual-based effect size may be viewed as a measure of the standardized group difference after controlling for the ability difference.

The uniform R^2 effect size is basically an increased portion of R^2 after entering the dummy language group variable into the ability-only regression model (Step 1); the non-uniform effect size, an increased portion of R^2 after adding the interaction term in the step-2 model. The total effect size is the aggregate of the uniform and non-uniform effects. Zumbo (1999) has suggested that, for an item to be classified as displaying DIF (*i.e.*, an aggregate of uniform and non-uniform DIF), the 2-degree of freedom χ^2 test between Step 1 and 3 have to have a p-value less than or equal to 0.01 and the R^2 difference between them should be at least 0.13 for the essay prompt. The Zumbo's (1999) DIF classification scheme of the R^2 values of 0.13 may correspond to a "medium" R^2 effect size in Cohen's (1988) standard. In Cohen's scheme, the R^2 effect sizes of 0.02, 0.13, and 0.26 are considered as "small", "medium", and "large" effect sizes, respectively, which can be also linked to the group mean score differences of 0.20, 0.50, and 0.80 in the standard deviation units. Thus, Zumbo's 0.13 threshold corresponds to Cohen's "medium" effect size and may be too high for some purposes.

The group-specific expected score curves were next obtained for those prompts that were flagged because of significant group effects as explained in the Appendix. For those prompts with significant ability-by-group interaction effects, the two separate group-specific curves cross at some point. For those prompts with no significant group effect, the curves are essentially identical. This can be regarded as a visual measure of the model-based effect sizes to show vividly the patterns of the uniform and non-uniform effects of the native language backgrounds on the essay scores. The vertical distance between the two lines at each ELA score point can be regarded as the expected essay score difference between the examinees of the same English language ability but from the different language groups.

Results & Discussion

English Language Ability versus Raw Essay Scores

Descriptive statistics of the 81 prompts used in the analysis are provided to give a general overview of the score information that was used for the logistic regression analysis for the two comparison groups. Table 1 reports overall means and standard deviations of the raw essay and the ELA scores for both the European and Asian language subgroups and standardized mean differences between the two groups, when 81 prompts were analyzed together. Figures 1-a and 1-b show the patterns of difference in the mean ELA and the mean raw essay score between the two groups across the 81 essay prompts.

As shown in Table 1 and Figures 1-a and 1-b, the ELA and raw essay scores were higher for the European language group than for the Asian language group both at the levels of the individual prompts and the total 81 prompts as a whole. At the level of the total prompts, the standardized mean difference in the ELA observed, .84, is quite significant and would be viewed as a "large" effect size (Cohen, 1988). The standardized mean essay score difference observed between the European and Asian language groups was 0.48, which is statistically significant and would be viewed as a "medium" effect size. Even at the level of each prompt, there was consistently a higher mean ELA score observed for the European language group across the 81 prompts. A similar pattern of difference was observed for the mean raw essay scores across the prompts. It would appear that the mean essay score difference between the two groups might be largely attributable to the difference in the mean ELA scores between the two groups overall. This difference is analogous to what is often called "impact" rather than "item bias" in the DIF literature (Clauser & Mazor, 1998).

A closer inspection of mean essay scores has revealed, however, that the distance between the two group score trend lines were less consistent across prompts in Figure 1-b than in Figure 1-a. On a few essay prompts, it appears that the differences in the mean raw essay scores are not explained by the differences in the mean ELA between the two groups alone. This suggests that the examinees' native languages might have some observable impact on essay scores as well.

Residual-Based Effect Sizes after Controlling for ELA

Mean expected essay scores from the step-one regression model (only with English ability included as a predictor) were averaged separately for each language group over the 81 prompts and are reported in Table 2. The mean residual scores for the two groups and the residual-based effect sizes for each of the 81 prompts are visually displayed in Figures 2-a and 2-b, respectively. The positive value of the residual-based effect sizes for a prompt indicates that the prompt is favoring the European group on average, whereas the negative indicates the other way around.

Mean expected scores over the 81 prompts were 4.23 and 3.79 for the European and the Asian groups, respectively, which were very close to the mean raw essay scores of 4.22 and 3.78. It may be that the English language ability differences have been already controlled for to a large extent by the CBT prompt selection algorithm that assigns a prompt to each examinee in a near-random way. Figure 2-a shows that the European language group performed better on some prompts but worse on others on average than expected for their English language ability, and so did the Asian language group. As shown in Table 2, the averaged residual-based effect size was almost zero when the data for all 81 prompts was combined because the negative and positive effect sizes were cancelled out across the prompts.

A closer examination of each prompt shows, however, that several prompts stand out from other prompts in the magnitude of the residual-based effect sizes, as shown in Figure 2-b. For instance, prompt 64 had the largest value of the mean residual-based effect size in magnitude (0.16), favoring the European language group as much as about 0.16 of the pooled standard deviation of the raw essay scores, while prompt 56 the second largest (about –0.14), favoring the Asian language group. Both of these two turned out to be less than "small effect sizes" according to Cohen's (1988) criterion. These prompts were flagged also in the 3-step modeling process that is described in more detail in the next section.

Slope Parameters and R^2 Effect Sizes from 3-Step Modeling

The results of the analysis based on the 3-step modeling process show that: (1) the English ability (x) variable was statistically by far the largest predictor of essay scores ($p<0.0001$) for all of the 81 prompts included in this analysis, as expected; (2) the group dummy variable (D_m) turned out to be a statistically significant predictor of the essay scores ($p=0.05$) for only 27 of the 81 prompts; (3) twenty of these 27 seven prompts exhibited a significant ability-by-group interaction ($x * D_m$) as well, but the remaining 7 prompts displayed the uniform group effect only (See Table 3).

Uniform R^2 effect sizes for the 27 flagged prompts ranged from to 0.0006 to 0.0073, with a mean of 0.0021 and a standard deviation of 0.0017. Shown in Table 4 are the 5 prompts with the largest uniform R^2 effect sizes and the largest slope parameters (β_2) for the dummy group variable (D_m) among the 27 prompts. The sizes of the absolute β_2 values were found to be proportional to those of the uniform R^2 effect. However, the

direction of the β_2 parameter value was positive for about a half of the prompts (14 prompts), but negative for another half (13 prompts). Prompt 64, for instance, had the largest positive slope values (0.42), favoring the European language to a small degree, whereas prompt 56 had the largest negative slope value (–0.34), favoring the Asian language group. All of the 5 prompts in Table 4 turned out to be the same ones that were identified as having the largest residual-based effect sizes in the previous section. Nevertheless, the sizes of the residual-based effect sizes correspond more closely to the total R^2 effect sizes rather than the uniform effect size alone.

Non-uniform effects were also exhibited in 20 of the 27 prompts with significant uniform group effect, with the non-uniform R^2 effect sizes varying from 0.0006 to 0.0032 among these 20 prompts. All of the remaining 7 prompts displaying uniform effect only had the positive β_2 values for the D_m variable, consistently favoring the European language group slightly at all ELA levels. Listed in Table 5 are the 5 prompts with the largest non-uniform effects sizes and the largest slope parameters (β_3) for the interaction term ($x * D_m$). The β_3 values for the interaction term were positive in direction for all of the 20 prompts, and the sizes of the β_3 values were also proportional to the non-uniform R^2 effect sizes. The non-uniform effect was larger than the uniform effect in only 11 of the 20 prompts, but smaller than the uniform effect in the remaining 9 prompts. For the eleven prompts with larger non-uniform effects, the non-uniform R^2 effect size ranged from 0.0010 to 0.0032.

Nevertheless, the increased R^2 values due to the group variable, the interaction variable, or both variables were far too small for any prompts to be regarded as displaying a serious level of uniform, non-uniform, or the combined DIF, when evaluated against Zumbo's (1999) criterion of 0.13, which seems a little high. Among the 5 prompts shown in Table 4, prompt 64 had the largest uniform and total R^2 effect sizes of 0.0073. On the other hand, prompt 67 had the largest non-uniform effect size of 0.0032 and the total effect size of 0.0038 among the 5 prompts listed in Table 5. All of these effect sizes are quite small by Cohen's criterion of a small R^2 effect size of 0.02.

Group-Specific Expected Essay Score Curves

To illustrate the direction and magnitude of the group effects visually, separate expected score curves were drawn for the European and Asian group for the 27 prompts that had the significant uniform group effect. In the case of the 20 prompts with the significant group main and the ability-group interaction effects together, the slope parameters for both the group (β_2) and the interaction variable (β_3) were entered along with the intercept (β_0) and the slope parameters (β_1) for the ELA variable into the equation for computing the group-specific expected essay scores and drawing the expected essay score curves for each group.

The directional and cross-over patterns of the group effect on each prompt are more clearly illustrated by drawing separate expected score curves for each language group, as shown in Figures 3-a and 3-b. On prompt 64, for example, the European language group is predicted to score higher than the Asian group consistently at all ELA score levels, because the slope parameter (β_2) for the dummy group variable (D_m) was

positive and the non-uniform group effect was not identified through the 3-step modeling process for this prompt. The vertical distance between the two score lines was the largest (0.18) at the ELA score point of about –7.4, whereas it was the smallest (0.08) at the ELA score point of 10, as shown in Figure 3-a. When the distances between the two curves were averaged across the actual ELA score range of about –10 to 5 in the data, the average distance was about 0.18, which is close to the residual-based effect size (0.16) previously computed for this prompt.

On prompt 56, on the other hand, the Asian language group tended to be advantaged considerably at the lower levels of ELA, but not at the higher levels, because the slope parameter (β_2) for the group variable was negative, with the slight non-uniform effect also present in the prompt. The vertical distance between the two expected score lines was the largest (-0.42) at the ELA score point of about –8.6, whereas the two curves crossed over at the ELA level of about 4.2. When the distances between the two curves were averaged across the actual ELA score range of –10 to 5 in the data, the average distance was about -0.22, which is a bit larger than the residual-based effect size (-0.14) in magnitude, but in the same direction, favoring the Asian language group overall.

A general pattern of expected scores for the prompt with both uniform and non-uniform effects was that the expected essay scores of the Asian group was higher at the lower ELA levels, but lower than, or equal to, those of the European language group at the higher ELA levels, as shown in Figure 3-b, 4-a, and 4-b. Nevertheless, the magnitude of the expected score difference varied from prompt to prompt, depending upon the direction and size of the slope parameter (β_2) values for the dummy group variable (D_m).

Summary and Conclusions

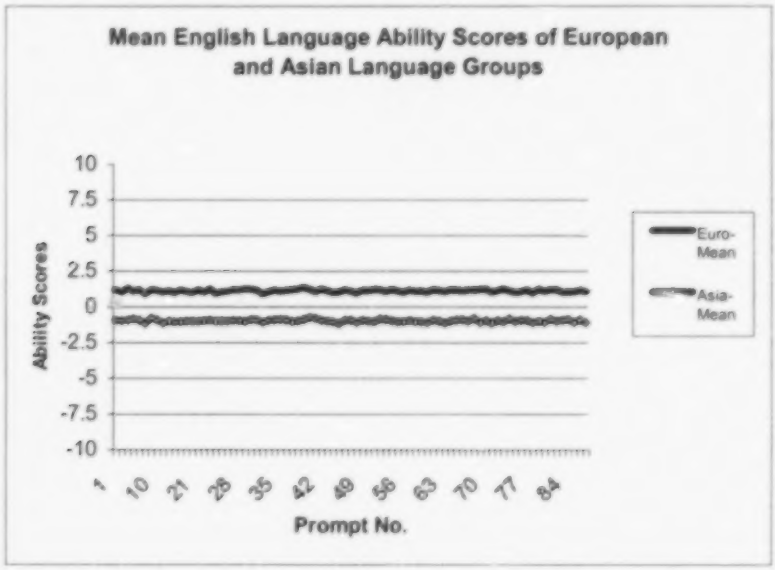
An extended version of the logistic regression DIF methodology employed in this study seemed to be efficient in investigating simultaneously both uniform and non-uniform group effect related to native languages. About two-thirds of the prompts analyzed in this study were found to be free of any native-language-related group effect. About one-third of the remaining prompts were initially flagged through the 3-step logistic regression method due to the significant uniform or non-uniform group effects, but their effect sizes were far too small for any of them to be classified as DIF essay prompts. In addition, the direction of the group effect was positive for some flagged prompts but negative for others and most of the flagged prompts have non-uniform effects as well. For this reason, the total effect size became almost zero at the aggregate level of the total 81 prompts analyzed in this study, because the positive and negative effect sizes were cancelled out across the essay prompts and across the ELA score levels for the prompts with the non-uniform effect.

Table 1
English Language Ability and Raw Essay Scores for
roups and Group Standardized Mean Differences

		Mean	S.D.	D
	494	4.22	0.95	0.48
	941	3.78	0.89	
	494	1.09	2.37	0.84
	941	-1.00	2.61	

Figure 1. Mean English Language Ability and Raw Essay Mean Scores of Reference Focal Groups for each of the 81 Essay Prompts

(a) Mean English Language Ability



(b) Mean Essay Scores

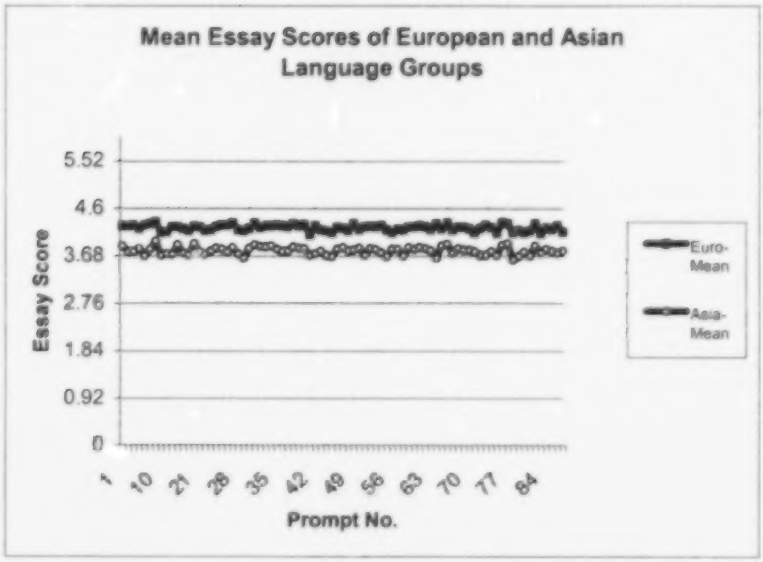
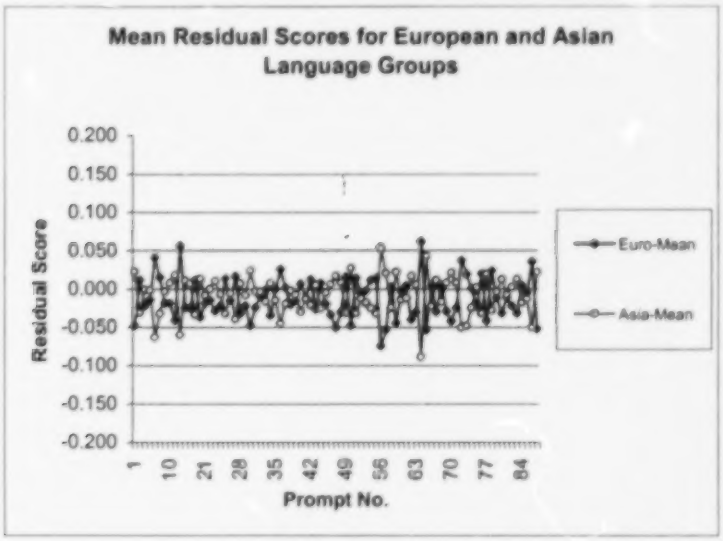


Table 2
Expected Mean Essay Scores, Residuals, and Standardized Mean Group Differences averaged Over 81 Prompts after Controlling for English Language Ability Differences Using Logistic Regression

Variable/Response Mode	Expected Score		Residual (Observed – Expected)		d
	M	SD	M	SD	
TOEFL Writing Score					
European Group	4.23	0.51	-0.01	0.77	0.00
Asian Group	3.79	0.55	-0.01	0.71	

Figure 2. Mean residual essay scores (Observed – Expected) of European and Asian language groups and the residual-based effect sizes for each of the 81 essay prompts

(a) Mean residual scores after controlling for English language ability



(b) Residual-based effect sizes

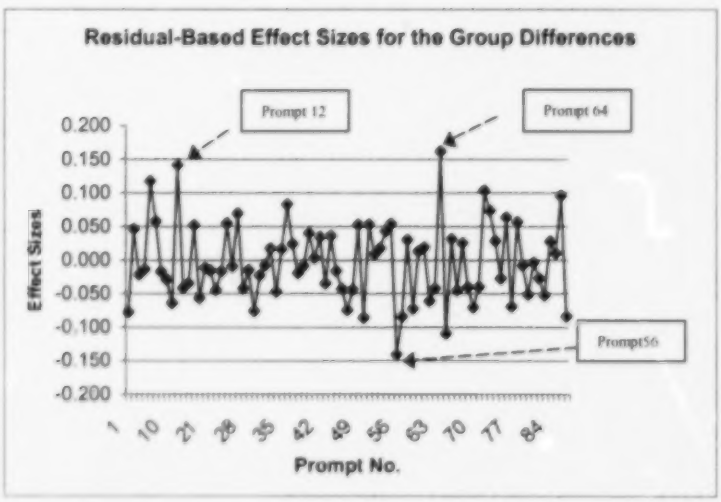


Table 3
Means of Slope Parameter and Increased R^2 value for the Added Predictor Variables in the Logistic Regression for the 81 prompts

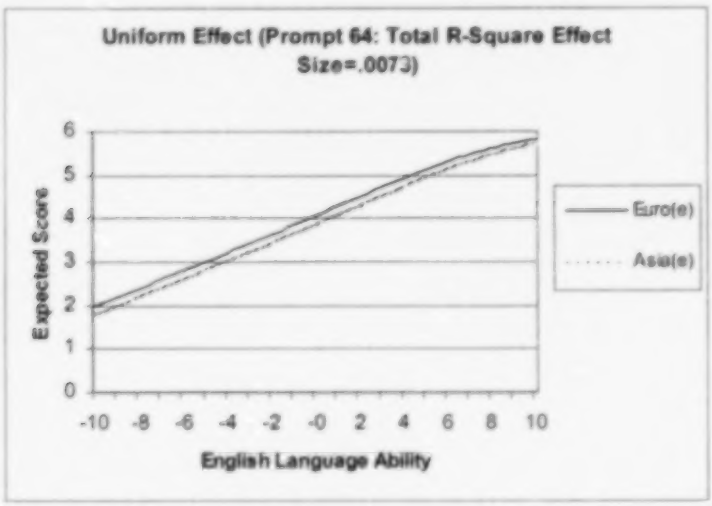
Group Effect		No. of Prompt	English Language Ability (x)		Native Language Groups (D_m)		Ability x Group Interaction ($x * D_m$)	
			Mean β_1	Mean R^2	Mean $\lvert\beta_2\rvert$	Mean R^2	Mean β_3	Mean R^2
No Group Effect		54	-0.5227	0.3766				
Uniform Only		7	-0.5055	0.3785	0.2420	0.0028		
Uniform + Non-Uniform	Uniform-Dominant	9	-0.5572	0.3717	0.2402	0.0024	0.0633	0.0012
	NU-Dominant	11	-0.5751	0.3780	0.1710	0.0011	0.0809	0.0019
Total		81	-0.5322	0.3764				

Table 4
Five Prompts with the Largest Uniform R-Squared Effect Size Estimated from the 3-Step Modeling Procedure

Prompt No.	No. of Examinees		Slope for Language Group (β_2)	R-Squared Effect Size		
	European	Asian		Uniform	Non-Uniform	Total
Prompt 64	1,357	1,434	0.4234	0.0073	0	0.0073
Prompt 12	1,087	1,231	0.3480	0.0051	0	0.0051
Prompt 56	1, 344	1,576	-0.3386	0.0043	0.0018	0.0061
Prompt 06	1,622	1,737	0.2745	0.0034	0.0015	0.0049
Prompt 65	1,750	1,772	-0.2803	0.0030	0.0011	0.0041

Figure 3. Separate Expected Score Curves for the Reference and Focal Groups Based on the Full Logistic Regression Model (Dominantly Uniform Effect)

(a) Uniform Effect Only Favoring the European Language Group (Prompt 64)



(b) Dominantly Uniform Effect Combined with Slight Non-Uniform Effect Mostly Favoring the Asian Language Group (Prompt 56)

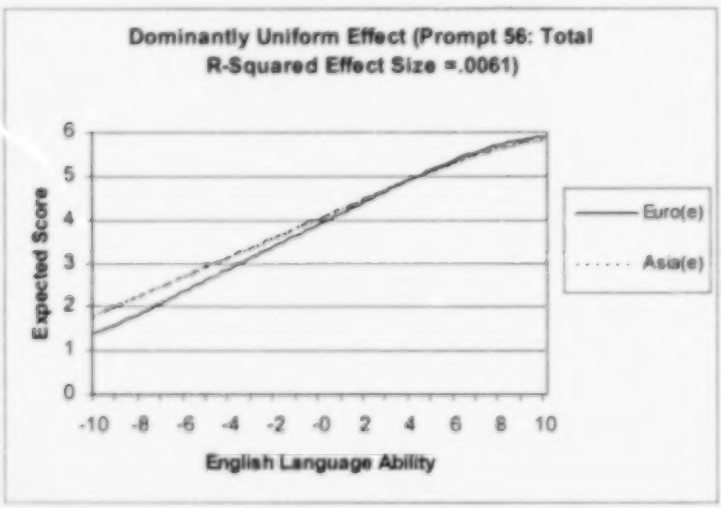
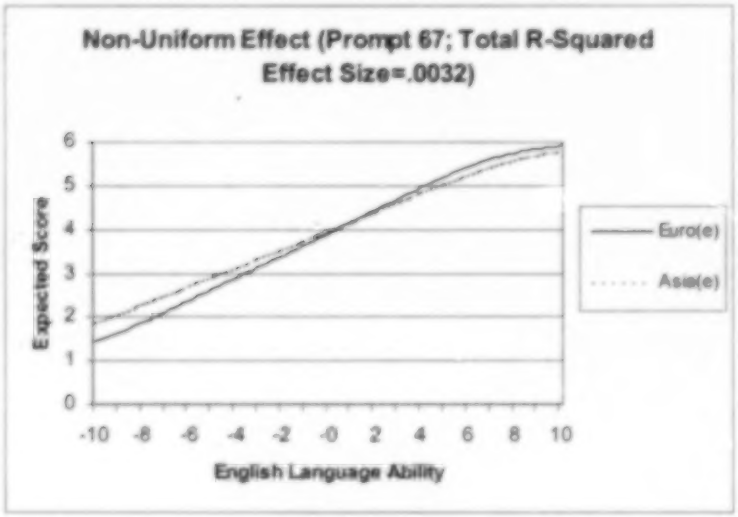


Table 5
Five Prompts with the Largest Slope Parameters for the Interaction Term and Non-Uniform R-Squared Effect Sizes from the 3-Step Modeling Procedure

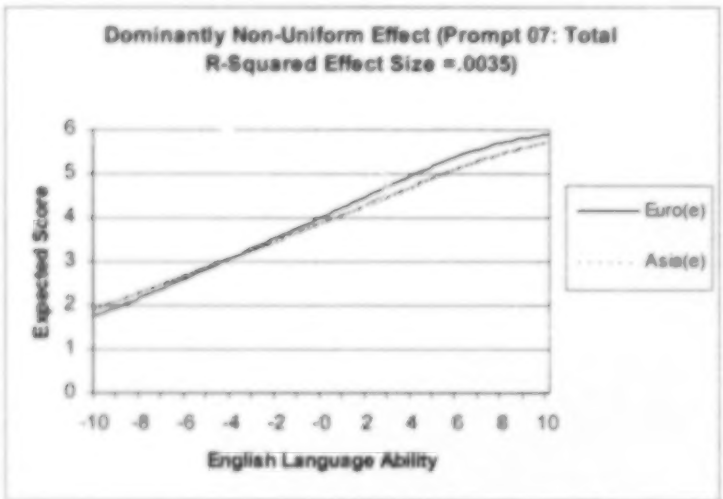
Prompt No.	No. of Examinees		Slope for Interaction Term (β_3)	R-Squared Effect Size		
	European	Asian		Non-Uniform	Uniform	Total
Prompt 67	1,729	1,799	0.1064	0.0032	0.0006	0.0038
Prompt 47	2,335	2,535	0.0964	0.0028	0.0012	0.0040
Prompt 07	1, 553	1,722	0.0948	0.0025	0.0010	0.0035
Prompt 57	1,789	2,060	0.0898	0.0022	0.0018	0.0040
Prompt 70	1,620	1,774	0.0875	0.0020	0.0013	0.0033

Figure 4. Separate Expected Score Curves for the Reference and Focal Groups Based on the Full Logistic Regression Model (Dominantly Non-Uniform Effect)

(a) Dominantly Non-Uniform Effects Combined with a Small Uniform Effect Favoring the Asian Language Group More, Especially at the Lower Ability Level (Prompt 67)



(b) Dominantly Non-Uniform Effects Combined with a Small Uniform Effect Favoring the European Language Group More Overall, Especially at the Higher Ability Level (Prompt 07)



APPENDIX: Derivation of the Logistic Regression Model

The multiple logistic regression equations for dichotomous items (i) can be written as:

$$P(U_i \mid x, D) = \frac{\exp[g_i(x,D)]}{1+\exp[g_i(x,D)]} = \frac{1}{1+\exp[-(g_i(x,D))]} \qquad (1)$$

where U_i represents the binary responses for dichotomized items i (U_i=0 or 1) and x is the continuous variable score, and D is the design matrix of the covariate variables. In this equation, the function $g_i(x,D)$ is called a *logit*. The logit is a linear combination of the continuous score (x), a covariate variable (D), and an interaction term (xD). If we want to analyze the DIF for M levels of a gender covariate, as in our TOEFL essay data, we can rewrite the logit $g_i(x,D)$ as:

$$g_i(x,D) = \beta_{0i} + \beta_1 x + \beta_2 D_m + \beta_3 x D_m \qquad (2)$$

where β_{0i} is the intercept for a dichotomous item (i), β_1 is the slope parameter associated with the English language ability score, β_2 is the parameter associated with the gender group variable, D_m, and β_3 is the slope parameter associated with the ability score-by-group interaction. In our study, D_m is 0 for male and 1 for female, respectively. It should be noted that the score-by-group interaction term was added to examine the score difference of non-uniform nature between the two groups.

The dichtomous model in Equation 1 can be directly extended for a polytomous item case based on the cumulative logit dichotomization scheme (Agresti, 1990; French & Miller, 1996). For the polytomous case, K+1 response categories for the polytomous item are dichotomized into K binary responses, and then the logistic regression is fitted to each dichotomized response for the ordinal item, with the parallel slopes assumed for all the dichotomized responses. In the actual TOEFL CBT essay data, there are eleven valid reported score categories (e.g., 1, 1.5, …, 5.5, 6), and, thus, there are 10 dichotomized responses (K-1). The proportional log-odds for each dichotomized response based on the cumulative logit scheme can be expressed as:

$$L_{jk} = \ln\left[\frac{\Pr(y_j \le k \mid x,D)}{1-\Pr(y_j \le k \mid x,D)}\right] = \ln\left[\frac{P_0(x,D)+P_1(x,D)+\ldots..+P_k(x,D)}{P_{k+1}(x,D)+P_{k+2}(x,D)+\ldots..+P_K(x,D)}\right] \qquad (3)$$

where L_{jk} stands for the proportional log-odds ratio for a dichotomized response (i) on the polytomous item (j), k is a subscript of the response category (k=0,1,2,….,K) for examinee scores (y) on the polytomous essay item, j. It should be noted that in this scheme the proportional log-odd ratio for this dichotomized response for prompt j is $\Pr(y_j \le k \mid x,D)$ over $[1-\Pr(y_j \le k+1 \mid x,D)]$, which is the opposite of Samejima's (1997) graded response model.

Category Characteristic Curves and Prompt Difficulty Index. If we define

$P_{jk}^{*}(x,D)$ and $P_{j,k+1}^{*}(x,D)$ as the regression of the binary item score method in which all score categories smaller than k and k+1, respectively, are scored 0 for each dichotomized item, the actual score category characteristic curve for score category k of the graded item j in relation to the independent variables x is

$$P_{jk}(x,D) = P_{j,k+1}^{*}(x,D) - P_{jk}^{*}(x,D) \qquad (4)$$

where

$$P_{jk}^{*}(x,D) = \sum_{v=0}^{k} P_{jv}(x,D)$$

Since the differencing scheme based on the cumulative logit logistic regression should be the opposite Samejima (1997), $P_{j0}^{*}(x,D)$ and $P_{j,K+1}^{*}(x,D)$ can be also defined in such a way that

$$P_{j0}^{*}(x,D) = 0$$

and

$$P_{j,K+1}^{*}(x,D) = 1$$

In the TOEFL CBT essay data, the score category response model for $y_j = k$ can be expressed by

$$P_{jk}(x,D) = \frac{\exp[(g_{j,k+1}(x,D)]}{1+\exp[(g_{j,k+1}(x,D)]} - \frac{\exp[(g_{jk}(x,D)]}{1+\exp[(g_{jk}(x,D)]} \qquad (5)$$